Soybean bacterial artificial chromosome contigs anchored with RFLPs: insights into genome duplication and gene clustering

Joann Mudge, Yan Huihuang, Roxanne L. Denny, Dana K. Howe, Dariush Danesh, Laura F. Marek, Ernie Retzel, Randy C. Shoemaker, and Nevin D. Young

Abstract: Surveying the soybean genome with 683 bacterial artificial chromosome (BAC) contiguous groups (contigs) anchored by restriction fragment length polymorphisms (RFLPs) enabled us to explore microsyntenic relationships among duplicated regions and also to examine the physical organization of hypomethylated (and presumably gene-rich) genomic regions. Numerous cases where nonhomologous RFLPs hybridized to common BAC clones indicated that RFLPs were physically clustered in soybean, apparently in less than 25% of the genome. By extension, we speculate that most of the genes are clustered in less than 275 M of the soybean genome. Approximately 40%–45% of this gene-rich portion is associated with the RFLP-anchored contigs described in this study. Similarities in genome organization among BAC contigs from duplicate genomic regions were also examined. Homoeologous BAC contigs often exhibited extensive microsynteny. Furthermore, paralogs recovered from duplicate contigs shared 86%–100% sequence identity.

Key words: Glycine max, bacterial artifical chromosome, restriction fragment length polymorphism, genome duplication, gene distribution.

Résumé : Un examen du génome du soja à l'aide de 683 contigs de chromosomes bactériens artificiels (BAC) marqués à l'aide de polymorphismes de longueur des fragments de restriction (RFLP) a permis aux auteurs d'explorer les relations de microsynténie parmi les régions dupliquées. Aussi, cela a permis d'examiner l'organisation physique des régions génomiques hypométhylées (vraisemblablement riches en gènes). L'observation de nombreux cas où des RFLP non-homologues ont hybridé aux mêmes clones BAC suggère que les RFLP sont groupés physiquement chez le soja, apparemment distribués sur moins de 25 % du génome. Par extension, les auteurs avancent que la plupart des gènes sont groupés au sein d'une portion du génome du soja qui correspondrait à moins de 275 Mpb. Environ 40 % à 45 % de cette portion du génome qui est riche en gènes serait représentée au sein des contigs décrits dans ce travail. Les similitudes dans l'organisation génomique au sein des contigs de BAC provenant de régions génomiques dupliquées ont également été examinées. Des contigs de BAC homéologues ont souvent montré beaucoup de microsynténie. De plus, les paralogues provenant de contigs dupliqués montraient 86 % à 100 % d'identité au niveau de la séquence.

Mots clés: Glycine max, chromosomes bactériens artificiels, polymorphisme de longueur des fragments de restriction, duplication génomique, distribution des gènes.

[Traduit par la Rédaction]

Introduction

The recently completed genome sequences of *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative 2000) and *Oryza sativa* (Goff et al. 2002; Yu et al. 2002) provide unprecedented and often unexpected insights into plant genome structure. Indeed, our general understanding of plant genome structure has improved dramatically in the last few years. As a model for plants, *Arabidopsis* was chosen in part because of its relatively small and (supposedly) simple genome (Pruitt and Meyerowitz 1986). Nevertheless, extensive duplicated segments were discovered in the completed *Arabi-*

Received 18 July 2003. Accepted 21 November 2003. Published on the NRC Research Press Web site at http://genome.nrc.ca on 24 March 2004.

Corresponding Editor: F. Belzile.

J. Mudge, Y. Huihuang, R.L. Denny, D.K. Howe, D. Danesh, and N.D. Young.[1] Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108, U.S.A.
L.F. Marek. Department of Agronomy, Iowa State University, Ames, IA 50011, U.S.A.
E. Retzel. Center for Computational Genomics and Bioinformatics, Academic Health Center, University of Minnesota, Minneapolis, MN 55455, U.S.A.
R.C. Shoemaker. USDA, Corn Insect and Crop Genetics Research Unit, and Department of Agronomy, Iowa State University, Ames, IA 50011, U.S.A.

[1] Corresponding author (e-mail: neviny@umn.edu).

doi: 10.1139/G03-141

dopsis genome (Blanc et al. 2000, 2003; Simillion et al. 2002; The *Arabidopsis* Genome Initiative 2000; Vision et al. 2000; Ziolkowski et al. 2003). The duplicated regions are derived from as many as three large-scale duplication events, altogether comprising nearly 80% of the *A. thaliana* genome (Simillion et al. 2002). The draft genomic sequence of rice also indicates a large amount of duplication, with 77% of genes having one or more duplicates and more than half of the sequences present in duplicated or triplicated genomic blocks (Goff et al. 2002).

Evidence for duplicated genomes has also been found in plant species whose full genomic sequences have not been completed. For example, homoeologous genomic segments in maize have been found to cover 60%–82% of the genome (Gaut 2001). A study of restriction fragment length polymorphism (RFLP) loci in soybean showed that each locus had on average 2.5 duplications in the genome. Adjacent RFLPs were often in the same set of duplicated regions, leading to duplicated chromosomal segments estimated to range from 1.5 to 106.4 cM in length (Shoemaker et al. 1996).

Evaluations of small-scale duplications in soybean are also available. Cross-hybridization and sequence similarity between a sample of paralogous genomic regions in soybean revealed highly similar duplicated regions (Foster-Hartnett et al. 2002). High levels of cross-hybridization were found in about half of the 15 sets of homologous contigs compared. Six sets of paralogous sequences from these groups showed sequence identity >98% even in cases where overall cross-hybridization was limited. Foster-Hartnett et al. (2002) also found cases in which duplicate contigs had nearly identical fingerprint patterns, distinguishable only on a sequence level.

Other studies have also found high levels of similarity between duplicate genomic regions in soybean. Yan et al. (2003) found that 86.5% of 37 duplicated bacterial artificial chromosome (BAC) contig sets exhibited microsynteny as measured by hybridization beyond that of the shared RFLP probe originally used to identify the duplicates. Sequencing of nine sets of paralogous sequences from one duplicated genomic segment showed 94% average identity over 3418 bp. Altogether, sequence similarity ranged from 81% to 96% (Yan et al. 2003). In further studies involving physical mapping and cross-hybridization of eight specific soybean contig groups, microsynteny in six cases, with extensive microsynteny in three, was observed (Yan et al. 2004). The majority of coding sequences were conserved (85%) and slightly fewer (75%) of low-copy (but apparently noncoding) sequences. Sequence order was found to be highly conserved when the three highly microsyntenic contig sets were examined in detail (Yan et al. 2004).

Researchers have speculated that genome duplication in plants generates sequence diversity, in contrast with extensive use of alternative splicing products (Yu et al. 2002). Indeed, polyploidy has long been known to be prevalent in plant species and to affect plant genome evolution (reviewed in Wendel 2000). Understanding genomic duplication in a species provides valuable insights into its evolutionary history, including its capacity to adapt and its genomic relationship with other species. On a practical level, duplicated genomic regions pose several challenges to genomic and genetic research, such as masking mutant phenotypes, complicating chromosome walking, and confounding the assembly of clones or sequences in shotgun approaches.

The physical layout of genes is also an important question in genomic research. If genes are clustered in the genome, researchers might be able to target gene-rich regions, effectively reducing the genome size that they must consider. The amount of gene clustering varies among plant species. In *Arabidopsis*, genes were found to be relatively evenly spaced throughout the genome (The *Arabidopsis* Genome Initiative 2000). Rice genomic sequencing also revealed largely even spacing of genes, although there were numerous regions of high gene density, especially on the distal ends of the chromosomes, interspersed with gene-poor areas (Goff et al. 2002; Yu et al. 2002).

In the absence of whole genome sequence data, gene distribution has been estimated by a variety of techniques. Hybridization of cDNA probes to density gradient fractions based on GC content predicted gene clustering in several large monocot genomes, including maize (Carels et al. 1995; Barakat et al. 1997) and barley (Barakat et al. 1997). Gene clustering was also found in dicot species, with the gene-containing regions of the genome expected to be 20% in pea, although tomato showed relatively little clustering (60%) (Barakat et al. 1999). In wheat, the distribution of gene-like sequences in deletion lines led to predictions that genes were clustered in less than 10% of the genome (Gill et al. 1996*a*, 1996*b*; Sandhu et al. 2001). Moreover, sequencing of large genomic segments has led to estimates of gene densities often approaching that of *Arabidopsis* (one gene per 4–5 kb; The *Arabidopsis* Genome Initiative 2000). For instance, Ku et al. (2000) sequenced a 105-kb BAC in tomato and found gene density to be one gene per 6.2 kb, only slightly lower than in *Arabidopsis* in spite of the sevenfold difference in genome size (Arumuganathan and Earle 1991). Panstruga et al. (1998) found gene density on a 60-kb segment of a barley BAC to be fivefold lower than in *Arabidopsis* but still 6- to 10-fold higher than expected if extrapolating from genome size.

Gene density estimates have also been obtained for legumes, including two model plants for the legume family, *Medicago truncatula* and *Lotus japonicus*. Gene density of *M. truncatula* has been estimated through BAC sequencing (D. Cook and D.J. Kim, University of California at Davis, and B.A. Roe, University of Oklahoma, Norman, Okla.). Results indicate gene density to be approximately one gene every 6.5 kb in the gene rich portion of the genome, estimated to cover one third of the genome. Cytogenetic studies on pachytene chromosomes of *M. truncatula* gave an even lower estimate of the gene space, producing estimates that euchromatic regions occupy only 20% of the genome, with the rest of the genome made up of gene-poor, repeat-rich heterochromatin (Kulikova et al. 2001). *Lotus japonicus* was found to have one gene every 9.9 kb based on the sequences of 183 TAC clones identified by gene-like sequences (Kaneko et al. 2003). With a genome size comparable with that of *M. truncatula*, this gives a higher than expected estimate of gene density when extrapolating from *A. thaliana* for genome size.

Soybean gene density has also been estimated. Foster-Hartnett et al. (2002) reported a gene density of one gene every 4.6 kb in the sequence of a 330-kb clone near *rhg1*, a gene for resistance to soybean cyst nematode. This estimate has since been revised to one gene every 8 kb (Young et al. 2003). Still, with a genome size almost nine times that of *Arabidopsis*, soybean's gene density is lower than *Arabidopsis* gene density by only a factor of 2.

More clues to plant gene organization will be forthcoming from the massive genome sequencing currently underway for many plants, including the legume models *M. truncatula* and *L. japonicus*. However, full genomic sequence data for the most economically important legume plant, soybean, is still probably several years away because of its larger and more repetitive genome. Nevertheless, genomic survey sequence and physical mapping can be used to increase our knowledge of genome structure and the implications for biological function and genome evolution.

Using BACs identified with low-copy RFLP probes in soybean, we have created an extensive network of several hundred anchored BAC contigs (contigs anchored by a single genetic marker and usually less than 200 kb in size) altogether covering more than 100 Mb of the soybean genome. Because the BACs were identified by hybridization to RFLPs, markers that generally hybridize to multiple regions of the soybean genome (Shoemaker et al. 1996), duplicate genomic regions were recovered and investigated. Our findings suggest that duplicate regions are highly conserved, with 86–100% sequence identity, and that interrelated sets of BAC contigs frequently exist, each comprising the same physically linked sets of RFLP markers. Because these RFLP probes were generated with a methylation-sensitive restriction enzyme, and are therefore derived from hypomethylated genome regions (Keim et al. 1990) overrepresentative of genic sequences (Burr et al. 1988), the distribution of genes could also be investigated. If RFLPs are clustered together in the genome, then the likelihood of RFLP loci being physically linked on the same BAC would occur more frequently than expected by chance. Based on this logic, our results suggest that RFLPs (and by extension, genes) are clustered in approximately 25% of the 1100-Mb soybean genome (Bennett and Leitch 2003). From this, we infer that the gene-containing regions of soybean occupy approximately 275 Mb, and contigs described in this study are estimated to cover up to 40%–45% of this gene-rich region.

Materials and methods

BAC identification

A base set of 304 soybean RFLPs were initially tested for potential use in identifying BAC minicontigs, defined as contigs averaging 180 kb in length anchored by an RFLP locus. These RFLPs were identified using the consensus soybean map (Cregan et al. 1999) and the "Soybase" database (http://soybase.ncgr.org/cgi-bin/ace/generic/search/soybase). The base set also included 16 additional RFLPs derived from BAC end clones and subclones, as well as a small number of disease resistance gene analogs (Peñuela et al. 2002) generated in the course of the project. Based on a preliminary analysis of these RFLPs, this set was reduced to a final total of 234 RFLPs. RFLPs found to have high copy number, RFLPs that identified no BAC clones, and RFLPs with questionable identities or high similarity to other RFLPs were excluded.

RFLP probes were radiolabeled and hybridized to BAC filters from two libraries as described previously (Marek et al. 2001). One BAC library was derived from *Hin*dIII partially digested genomic DNA of the variety "Williams 82" (Marek and Shoemaker 1997) and the other from *Eco*RI partially digested genomic DNA of the variety "Faribault" (Danesh et al. 1998). Each filter consisted of more than 18 000 BAC colonies, spotted in duplicate, so a complete hybridization experiment required a total of six high-density filters. DNA of up to 16 representative BACs was typically miniprepped (alternative method of Marra et al. 1997) for each probe to build BAC contigs and obtain end sequence, as described below.

Fingerprinting analysis

Because the soybean genome is highly duplicated (Shoemaker et al. 1996; Marek et al. 2001), most RFLP probes identified more than one genomic region. Thus, BACs identified with a single RFLP probe had to be grouped into physically distinct contigs as described in Marek et al. (2001). Briefly, Southern blots were prepared with DNA from identified BACs along with genomic DNA from the parents of the mapping population(s) originally used to map the corresponding RFLP locus. All DNA samples were digested with the enzyme used to map that RFLP locus. In this step, a maximum of 16 positive BACs (usually those with the strongest signals) were typically analyzed from each library hybridization, even if additional positive clones had been observed. Restriction fragment patterns obtained from hybridization with the radiolabeled RFLP probe were then used to group the BACs into distinct contigs, and the contig with a restriction fragment matching that of the mapped RFLP locus was tentatively assigned the map location of the underlying RFLP locus. When multiple contigs were identified by a single RFLP, the contigs were named contig_a, contig_b, and so on in arbitrary order.

Many of the contigs were analyzed further to confirm contig assignment through the use of agarose-based *Eco*RI fingerprint analysis, as described in Marek et al. (2001). In addition, BAC clones that had not been chosen for the initial fingerprint analysis, but that were subsequently identified by a different second probe that identified an overlapping set of BACs (see below), were also included in the *Eco*RI fingerprinting analysis. Contigs were reassembled from fingerprints using the programs Image (version 3.10) (Sulston et al. 1988) and FPC (version 4.7.9) (Soderlund et al. 1997) with a cutoff of $p < 0.00001$. This relatively lenient value was reasonable, as the underlying BAC contigs had previously been assigned to (putative) contigs based on the Southern analysis described above.

BAC end sequencing and sequence analysis

BACs were sequenced at both ends using the M13 forward primer and a modified M13 reverse primer (5′-TTC ACA.CAG GAA ACA GC-3′). Sequencing was performed at the University of Minnesota Advanced Genetic Analysis Center on ABI PRISM 377 sequencers using BigDye terminator cycle sequencing kits (Perkin Elmer). Sequence trace files were processed at the University of Minnesota's Center for Computational Genomics and Bioinformatics. Base calling, performed with PHRED, required a minimum quality score of 10 (Ewing and Green 1998). Additional quality control measures were also taken for each sequence as follows. Sequences were trimmed to limit the number of unknown bases to 7%, vector sequences were removed (Shoop et al. 1995), and sequences were further trimmed to ≤4% unknown bases. Sequences greater than 200 bp were submitted to the Genome Survey Sequence Database (dbGSS) at GenBank. Altogether, a total of 3344 high-quality BAC end sequences, averaging 470 bp, were generated (over 1.5 Mb in total).

Isolation of DNA paralogs in soybean

To better understand the divergence of homoeologous regions in the soybean genome, primer pairs were designed from several BAC end sequences for use in amplifying DNA from BACs in other contigs that were confirmed by Southern hybridization to be located in a duplicated genomic region. Miniprepped DNA from BACs in duplicated contigs that showed strong cross-hybridization to these BAC end probes was used as template. PCR products were either sequenced directly or cloned into the pGEM®-T Easy Vector (Promega, Madison, Wis.) before being sequenced at the University of Minnesota Advanced Genetic Analysis Center. These PCR products, as well as the 14 BAC ends used to identify primer pairs, were sequenced at least twice, and sequence alignments were performed with Sequencher-TM3.1.1 (Gene Codes Co., Ann Arbor, Mich.). Sequences were designated as genic if they had Blast hits (Altschul et al. 1997) against the “Peptides” database (BLASTx; expect value $\leq 1 \times 10^{-8}$), a database housed at the Center for Computational Genomics and Bioinformatics and containing nonredundant protein sequences from SwissProt and TrEMBL (Bairoch andApweiler 2000), PIR and NRL3D (Barker et al. 2000), and GenPept (Benson et al. 2003). Sequences were also designated as genic if they had Blast hits to soybean expressed sequence tags from dbEST (Benson et al. 2003; Blastn, percent similarity ≥95% and alignment length ≥100 bp).

Physical linkage of RFLP loci

RFLP loci were considered physically linked and their associated contigs “coalesced” by looking for probes that identified common sets of BAC clones. Putative coalesced contigs were assigned as follows. (*i*) Contigs from two different probes were coalesced if at they shared at least two BACs in common. (*ii*) If only one BAC was held in common, contigs were coalesced if either (*a*) another pair of contigs from the same set of probes had been coalesced because they shared two or more BACs or (*b*) the probes shared one or more additional BACs in common that had not been previously assigned to contigs.

DNA fingerprinting was used to confirm many of the coalesced BAC contigs and BAC filters were rechecked to ensure that BACs that were positive with only one of the two probes had not been overlooked when the filters were initially read. Representatives from all contig groups identified with a given RFLP probe were included in this stage of the analysis. Further quality control was conducted by eliminating all pairs of RFLP probes with sequence alignments of 80% or more over at least 100 bp or that showed significant similarity to any repetitive sequence.

Distribution of RFLP loci

The distribution of the number of redundant versus unique BACs identified was analyzed. Redundant BACs were defined as those BACs that had previously been found with another RFLP probe, whereas unique BACs were those found for the first time. The distribution of redundant versus unique BAC hits was compared with a theoretical distribution to determine if RFLPs sample BACs from the soybean genome in a biased (nonuniform) manner. A sample of 5532 BAC hits from 195 RFLP probes was examined. Expected numbers of redundant versus unique BAC hits when sampling from the 110 592 BACs in the libraries were simulated with a Poisson distribution. In addition, the possibility that RFLPs sample the genome nonuniformly was tested by comparing the actual distribution with expected numbers obtained from Poisson distributions simulating sampling of different fractions of the libraries at 1% intervals. Observed and expected numbers of redundant and unique BAC hits were compared using a χ^2 test.

RFLPs derived from BAC end clones and subclones, or resistance gene analogues were excluded in this analysis. RFLP probes positive with repetitive DNA and redundant probes defined as those having 80% identity over 100 bp or more with another probe in the analysis were also excluded. Finally, BACs that were positive on the initial library hybridizations but not on subsequent fingerprint analysis probed with the corresponding RFLP were excluded.

Results

Hybridization of two soybean BAC libraries with 234 RFLPs produced a network of 683 total BAC contigs. On average, contigs comprised 3.4 BAC clones after analyzing approximately one third of the positive BACs observed and averaged 180 kb in length. Of the contigs, 240 could be tentatively anchored to the soybean genetic map (Cregan et al. 1999) by matching the size(s) of one or more of their restriction fragments to the underlying RFLP marker or through the use of simple sequence repeats, single-nucleotide polymorphisms, or cleaved amplified polymorphic sequences found in BAC end sequences. Because RFLP probes identified an average of 2.9 genomic regions (data not shown), these contigs corresponded to over 200 contig sets, groups of contigs related by a common RFLP marker. A total of 3344 end sequences were obtained from BACs and, in a few cases, from subclones. Some BAC end sequences were used to identify and examine paralogs from duplicate contigs.

Divergence of paralogs in duplicate genomic regions

To identify paralogous sequences, PCR primers were developed from several BAC end sequences. A small set of 18 primer pairs that amplified the BAC end sequences were examined in detail. Four did not amplify the putatively paralogous regions in duplicate contigs, even though the original BAC end did cross-hybridize to the duplicate contig. The 14 primer pairs from 10 contig groups that did amplify duplicate contigs were used to sequence paralogous regions. Each primer pair was used to amplify paralogous se-

Table 1. Sequence comparison between paralogous sequences.

	Original contig				Duplicate contig					
RFLP probe	Contig	BAC end	GenBank accession No.	Genic[a]	Contig	BAC	Deletion (bp)	Mutation (bp)	Length (bp)	Identity (%)
pA064	a	Gm_UMb001_185_I06R	AQ936136	No	d	Gm_UMb001_024_J18		12	229	94.8
pA064	d	Gm_UMb001_024_J18F	AQ936100	Yes	a	Gm_UMb001_185_I06		10	203	95.1
pA112	a	Gm_UMb001_030_H10F	AZ045086	Yes	b	Gm_UMb001_059_P24			338	100.0
PA112	a	Gm_UMb001_030_H10F	AZ045086	Yes	c	Gm_UMb001_124_M03		4	338	98.8
pA112	a	Gm_ISb001_002_A03R	AZ045123	No	b	Gm_UMb001_059_P24	1	2	391	99.2
pA112	a	Gm_ISb001_002_A03R	AZ045123	No	c	Gm_UMb001_124_M03		2	394	99.2
pA112	a	Gm_ISb001_002_A03R	AZ045123	No	d	Gm_UMb001_159_B15	3		395	99.2
pA257	b	Gm_UMb001_093_I19R	AZ045036	No	a	Gm_UMb001_037_H05	7	13	287	93.0
pA401	b	Gm_Umb001_157_G01F	AQ810543	No	d	Gm_UMb001_131_K24	25	37	464	86.3
pA401	c	Gm_Umb001_110_J04F	AQ810545	Yes	a	Gm_UMb001_129_D20	1	6	166	95.8
pA401	c	Gm_Umb001_110_J04F	AQ810545	Yes	b	Gm_UMb001_157_G01	1		202	99.5
pA702	b	Gm_ISb001_046_H02F	AQ936114	Yes	a	Gm_UMb001_107_F18			338	100.0
pA810	b	Gm_UMb001_139_N10F	AZ045025	Yes	a	Gm_ISb001_003_H22	4	13	242	93.0
pA885	a	Gm_Umb001_015_P01F	AQ852244	Yes	b	Gm_UMb001_029_E22			220	100.0
pB208	b	Gm_ISb001_023_N23F	AQ936041	Yes	a	Gm_ISb001_033_G01	4	12	225	92.9
pB208	b	Gm_ISb001_023_N23R	AQ936046	Yes	a	Gm_UMb001_021_K21	1	13	247	94.3
Bng171	a	Gm_UMb001_021_D05F	AQ851471	No	b	Gm_UMb001_093_M12	11	12	206	88.8
26_J06R	a	Gm_UMb001_008_A22R	AZ044731	No	c	Gm_UMb001_001_A05	1	3	227	98.2
Total							59	139	5112	96.12

[a] Coding sequence (1 × 10^{-8} in BLASTx or 95% similarity over 100 bp in BLASTn to soybean expressed sequence tag).

quence(s) on BACs in one to three of its duplicate contigs. This resulted in 18 paralogous sequences that could be compared (Table 1).

Paralogous sequence amplification yielded PCR products with an average of 96.1% identity to the original BAC end sequences. Overall, similarity between paralogous sequences ranged from 86.3% to 100% (Table 1). Divergence resulted from a combination of insertion/deletion events (1.2%) and point mutations (2.7%). Surprisingly, identity was slightly lower in alignments of genic sequences compared with alignments of sequences without evidence of being genic (95.2% versus 99.0%). In one case (pA112), four related contigs were examined for sequence identity through five sequence comparisons. Unexpectedly, all of the comparisons among paralogs yielded similar levels of sequence identity (98.8%–100%).

Physical linkage of RFLP loci

The 683 BAC contigs initially identified with RFLP hybridization were examined in further detail to uncover contigs containing two or more physically linked, yet distinct (nonhomologous), RFLP loci. Contigs associated with linked RFLP loci were tentatively “coalesced”.

When coalescing contigs, there were eight cases in which two distinct contigs, both from the same probe, were merged with a single contig from another probe. Contigs were created based on restriction fragment patterns. It is possible that two distinct contigs identified with a probe had restriction fragment patterns that were highly similar and therefore were incorrectly merged into a single contig. This would result in the overlap of two contigs from one probe with one incorrect contig from another. Alternatively, two contigs may have been incorrectly formed from a single contig. No further work was done to determine whether there should have been one or two contigs in these cases.

Based on criteria described in the Materials and methods, of the 683 BAC contigs identified with RFLPs, 76 sets of contigs could be coalesced on the basis of shared BAC clones and physically linked RFLP loci. Because several probes had more than one coalescing contig and because three or more contigs from different probes often coalesced together, the coalesced contigs could be sorted into 28 independent groups. Each group contained two to nine probes forming a network of physically linked RFLP loci. In some cases, all of the probes in a network were physically linked to one another in one genomic region. In other cases, subsets of loci were physically linked in duplicate genomic regions.

At least 16 of these groups showed physical linkage of the same pairs of RFLP markers in duplicate regions of the genome. For instance, Bng173 and pA975 were physically linked on two different homoeologous contigs (Fig. 1). One of the contigs mapped to the top part of linkage group G (Cregan et al. 1999) near *rhg1*, a gene for soybean cyst nematode resistance (Foster-Hartnett et al. 2002).

Some RFLP loci that show linkage due to the overlap of BACs that they have identified also show linkage on the genetic map. For instance, RFLP probes pA343 and pB142 are linked on a contig and cosegregate on genetic linkage group D (Shoemaker and Olson 1993; http://soybase.ncgr.org/cgi-bin/ace/generic/search/soybase) (Fig. 2A). Each has a homoeologous contig on the composite map on linkage group B2 (http://soybase.ncgr.org/cgi-bin/ace/generic/search/ soybase), although these do not overlap (Fig. 2B). They are separated by 3.1 cM but do not show physical linkage in this study.

Many groups of coalesced contigs show linkage of several different RFLP probes across duplicated regions, resulting in a complex network of overlapping and duplicate contigs. For instance, pA256 anchored three contigs in this study (Fig. 3). Two of its loci have been placed on the composite map on linkage group A2 at 0.8 cM (A256_1) and on linkage group A1 at 76.7 cM (A256_2). One of the contigs in this study, Gm_A256_ctg_a, has been shown to match A256_2 on linkage group A1. Each of the three pA256 contigs coalesced with contigs identified by other probes. Although they are duplicates of each other, each showed a different pattern of coalescing. Gm_A256_ctg_b coalesced with Gm_B174_ctg_b (Fig. 3A). Contig Gm_A256_ctg_a coalesced with Gm_A381_ctg_c (Fig. 3B). In addition, contigs Gm_A256_ctg_a and Gm_A256_ctg_c each coalesced with a contig from pK636, contigs Gm_K636_ctg_e and Gm_K636_ctg_d, respectively (Figs. 3B and 3C). Although no experimental information is available about where the pK636 contigs map, two loci on the composite genetic map are on A2 at 77.2 cM (K636_1) and on A1 at 0 cM (K636_2). The overlap between Gm_K636_ctg_e and Gm_A256_ctg_a implies that the former matches the genetic locus K636_2, which is only 0.5 cM away from locus A256_2 and Gm_A256_ctg_a. Furthermore, it is possible that the other two matching contigs from these probes, Gm_A256_ctg_c and Gm_K636_ctg_d, match A256_1 and K636_1, respectively, which are 0.8 cM apart on the composite genetic map.

RFLP locus distribution

The distribution of BACs uniquely identified by a single RFLP probe versus those that had already been identified by one or more other RFLP probes was compared with expected numbers obtained from a Poisson distribution. To test whether RFLP probes are not randomly distributed throughout the genome, and hence sample only a portion of it, distributions were also simulated for partial sampling of the library, simulating partial sampling of the genome.

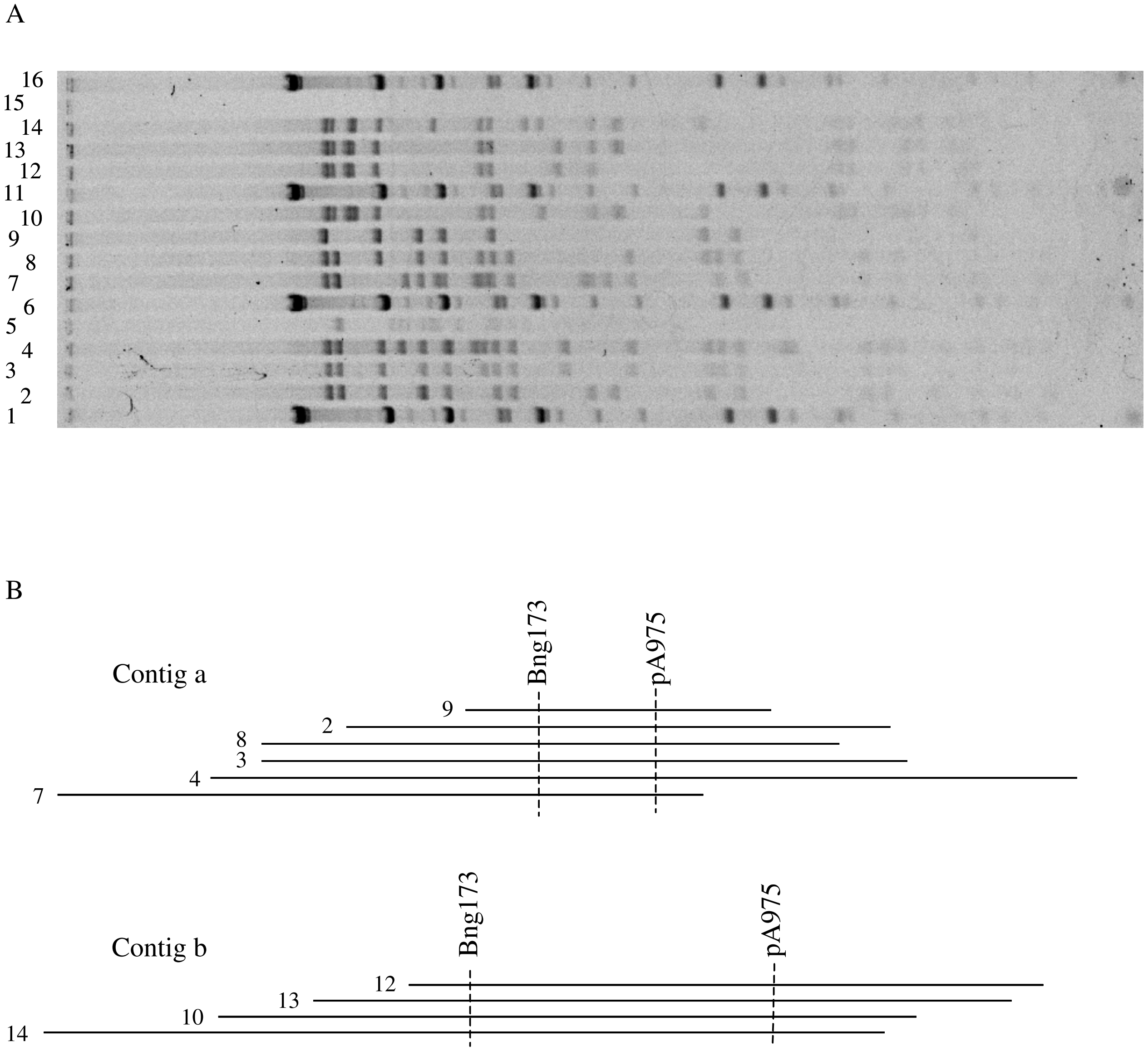
The distribution of unique versus redundant BAC identification was significantly different from the expected distribution when randomly sampling all BACs ($p = 1 \times 10^{-277}$) (Fig. 4). In fact, there were approximately four times more redundant BAC hits found in this project than would be expected when randomly sampling from our set of 110 592 BACs. However, the observed distribution did not differ significantly from that expected if sampling 22%–25% of the BACs in the library (Fig. 4). The most likely percentage of the library that the RFLP probes were sampling was 24% ($p = 0.73$). This model, sampling from 24% of the genome, predicts 538 redundant hits, close to the observed number of 546. For comparison, the model of sampling the entire genome (no clustering of RFLP probes) predicted just 136 redundant BAC hits ($p = 1 \times 10^{-277}$).

Discussion

Duplicate genomic regions

Extensive networks of duplication within the soybean ge-

Fig. 1. Contigs containing two RFLP probes, Bng173 and pA975. (A) *Eco*RI digestion of BACs identified by these two probes. Lanes 1, 6, 11, and 16 contained molecular weight marker (Marra et al. 1997). Lanes 2–5 and 7–9 contained DNA from BACs in Bng173 and pA973 contigs a. Lanes 10 and 12–14 contained DNA from BACs in Bng173 and pA973 contigs b. No DNA was loaded into lane 15. (B) Diagram of contigs generated by FPC (Soderlund et al. 1997). Numbers correspond to the lane numbers in Fig. 1A. Positions of the RFLP probes, marked with broken lines, are arbitrary.

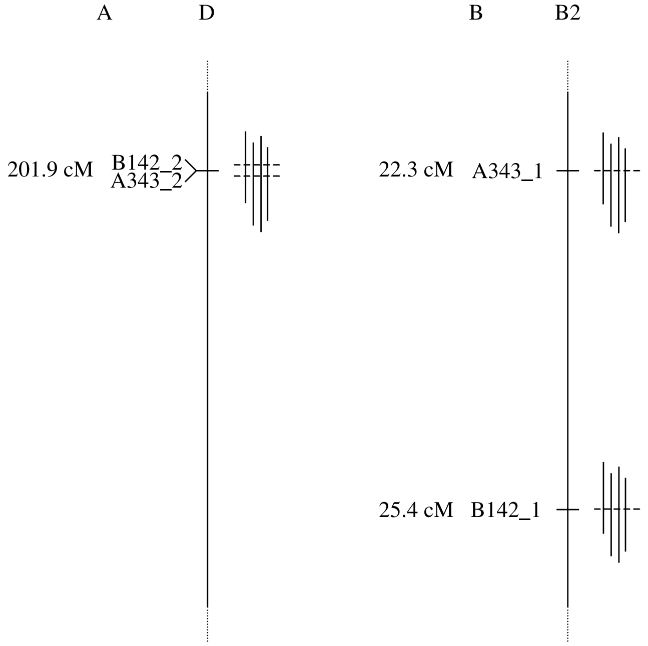


nome have been uncovered in this study. Over 200 RFLP probes have been used to anchor an average of 2.9 genomic regions each, similar to the estimate of 2.5 genomic loci per RFLP probe found by Shoemaker et al. (1996). Comparison of these duplicate genomic regions on a kilobase pair scale has revealed several different lines of evidence supporting a high degree of similarity between duplicated regions.

Sequencing of paralogous regions from homoeologous contigs showed that duplicated regions are highly similar on a sequence level. Eighteen sets of paralogous sequences from 10 different duplicate contig sets exhibited 86%–100% identity in sequence. The 18 paralogous sequences, containing more than 5 kb altogether, averaged 96% identity with the original sequences. However, 16% of the primer pairs in this study failed to amplify the (putatively) paralogous sequence in spite of the fact that hybridization showed the presence of a paralogous sequence. Sequence divergence in the primer region was probably responsible for the lack of amplification in these cases. Indeed, if forward and reverse primers of 18 bp are generated from these regions, assuming an average of 96% identity, then 5% of them will be expected to have less than 90% identity with their paralagous sequence.

Paralogous sequences gave us another glimpse at the similarity between duplicate regions by allowing us to observe microsynteny between duplicate contigs. Each contig set is known to be related by hybridization to a common RFLP probe. The paralogous RFLP loci together with the sequencing of another paralogous set of sequences for each group (Table 1) indicates microsynteny, or physical linkage of paralogs. In addition, three groups have enough data to indicate more extensive levels of microsynteny. Contigs a and d of probe pA064 each contain sequences paralogous to RFLP

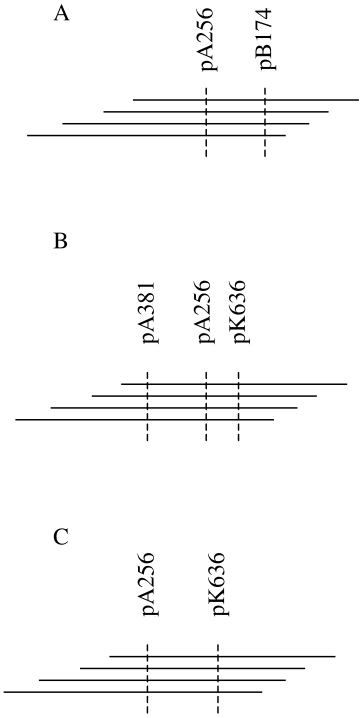
Fig. 2. Schematic view of two duplicate regions anchored by pB142 and pA343. Each RFLP probe has a locus on (A) linkage group D (Shoemaker and Olson 1993) and on (B) the composite linkage group B2 (http://soybase.ncgr.org/cgi-bin/ace/generic/search/soybase). The contigs on linkage group D overlap while those on linkage group B2 do not.



probe pA064 as well as BAC end sequences Gm_UMb001_185_I06R and Gm_UMb001_024_J18F. Contigs a and b from probe pB208 also show physical linkage of three paralogs in each contig: paralogs of RFLP probe pB208 plus both ends of BAC GM_ISb001_023_N23. In the pA112 contig set, contigs a, b, and c all show conserved linkage of three sequences, related to RFLP probe pA112 and BAC end sequences Gm_UMb001_H10F and Gm_ISb001_002_A03R. Microsynteny between duplicated and even triplicated regions provides additional confirmation of the high degree of similarity between paralogs.

More evidence of similarity between duplicate genomic regions is provided by the physical linkage of two or more gene-like RFLP loci in duplicate genomic regions. In more than half of the coalesced contig groups, duplicate contigs shared the same physically linked RFLP markers. For instance, RFLP probes Bng173 and pA975 have physically linked loci in two duplicated regions of the genome. Loci from pA256 and pK636 are also linked in two different regions of the genome, although the different patterns of linked RFLP loci on the three pA256 contigs also highlight the complex relationships between duplicated regions. Even RFLP probes pA343 and pB142, which this study shows as physically linked in only one genomic region (on linkage group D1b+W where they are known to have genetic loci within 1 cM of each other), have duplicate loci on linkage group B2 separated by 2.1 cM. This genetic distance is presumably too far away to be detected in this study, although there is physical linkage in duplicated genomic regions. With many of the duplicated locations of RFLP loci unknown, we are probably only capturing a portion of physically linked RFLP probes, missing many that, although physically linked, are at a distance beyond what can be detected here.

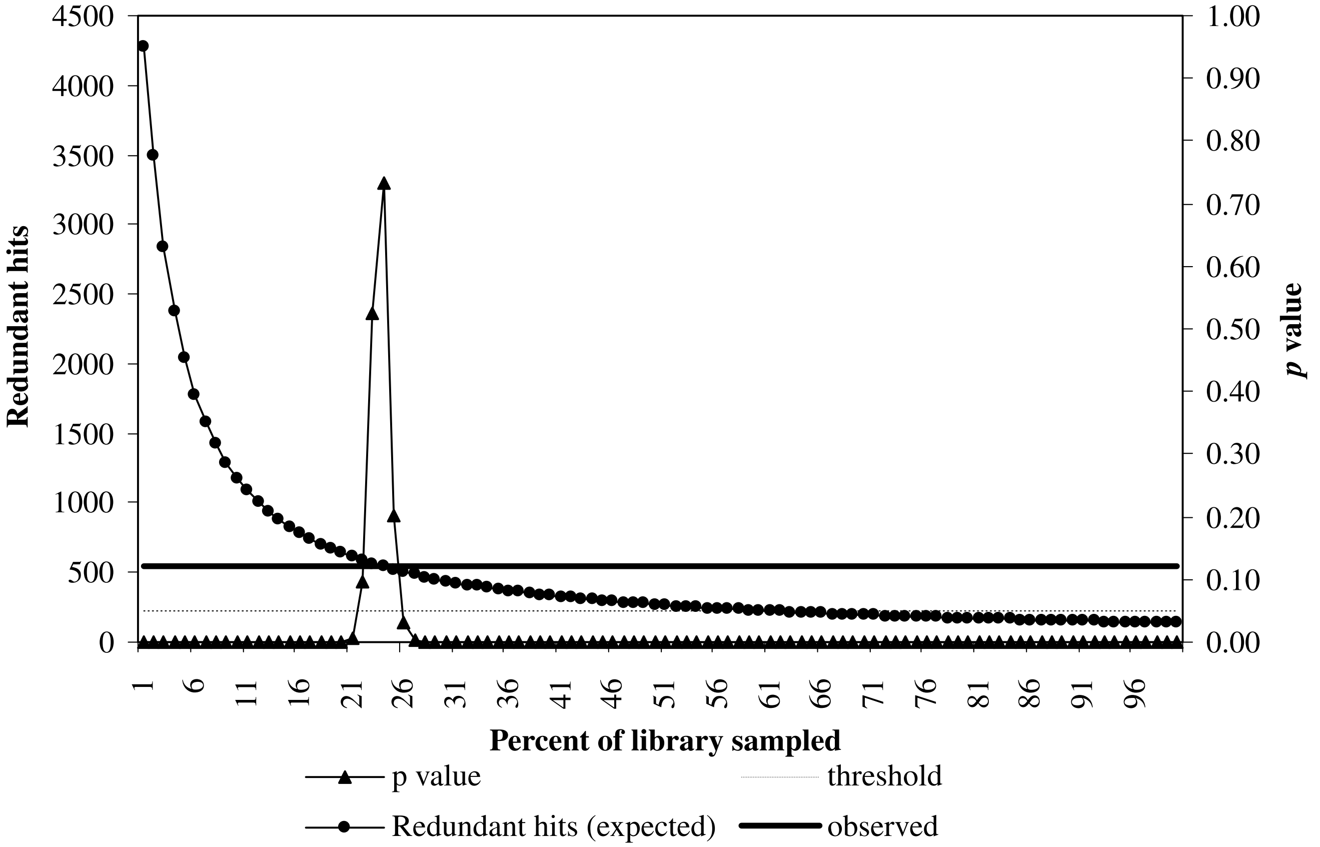
Fig. 3. Schematic of contigs from a triplicated genomic segment anchored by probe pA256. Placement of the RFLP loci is arbitrary. (A) Coalesced contig from contigs Gm_A256_ctg_b (known to be anchored by the RFLP locus A256_2 on linkage group A1) and Gm_B174_ctg_b having uncharacterized linked loci of pA256 and pB174. (B) Contig Gm_A256_ctg_a was coalesced with Gm_A381_ctg_c and Gm_K636_ctg_e. The resulting contig shows physical linkage of the RFLP locus, A256_2, known to map to linkage group A1, and uncharacterized loci of pA381 and pK636. Gm_A381_ctg_c and Gm_K636_ctg_e did not show overlap with one another. (C) Uncharacterized loci of pA256 and pK636 were also linked in a related genomic region where their two contigs, Gm_A256_ctg_c and Gm_K636_ctg_d, were coalesced. Positions of the RFLP probes, marked with broken lines, are arbitrary.



As more is learned about these duplicate genomic regions, more synteny of paralogs is likely to be observed. For instance, hints of the redundancy and shared ancestry between pA975 and Bng173 are available on Soybase (http://soybase.ncgr.org/cgi-bin/ace/generic/search/soybase), although none of their mapped loci coincide. Bng173_1 is only 2 cM away from a gene for resistance to soybean cyst nematode (SCN) on the composite map of linkage group G. Two of the three loci of A975 mapped on the composite map are in the vicinity of SCN loci. A975_4 overlaps an SCN locus on linkage group E and A975_2 is less than 3 cM from an SCN locus on linkage group A2. Further similarity between the A975 duplicated loci can be discovered on Soybase in that all three loci are in the vicinity of *Sclerotinia* disease resistance genes.

Finally, coalescing contigs based on physical linkage of RFLP loci also highlighted the similarity between duplicate regions by identifying duplicate contigs with indistinguish-

Fig. 4. Comparison of the observed and expected numbers of redundant BAC hits. The expectations were developed based on the "sampling efficiency" of the RFLP probes, that is, the percentage of BACs in the library that they sample. The line with circles marks the number of redundant BAC hits expected at different RFLP sampling efficiencies. The solid horizontal line marks the observed number of redundant BAC hits (546). A total of 5532 BACs were sampled and this number was also used for calculating the expectations, which were based on Poisson distributions as described in the text. The line with triangles tracks the *p* value (right-hand axis) for the χ^2 test of the null hypothesis that the expected numbers of unique and redundant BAC hits are not significantly different from the observed distribution. The broken horizontal line marks the threshold for determining significance (p = 0.05).



able restriction fragment patterns. When examining overlapping contigs, there were eight cases in which one contig shared BACs with two distinct contigs from another probe. This difficulty apparently arose from incorrect merging of distinct contigs from highly similar duplicated regions or splitting of a single contig. This emphasizes the difficulty in constructing BAC contigs based on restriction fragment patterns in a highly duplicated genome such as soybean.

These findings of highly similar duplicated and triplicated genomic regions confirm earlier studies of microscale similarity in soybean. Looking at eight genetically linked BAC contigs and one or more duplicates for each, Foster-Hartnett et al. (2002) compared homoeologous regions using cross-hybridization, sequencing, and fingerprint patterns. Over half of the contigs showed extensive cross-hybridization. Of six comparisons of paralogous sequences between original and duplicate contigs, all showed at least 98% identity, including one that had shown only low levels of cross-hybridization. That same study also found high levels of similarity as shown by identical fragments in restriction patterns in more than half of the BAC contig sets examined.

In comparing 37 sets of homoeologous contigs, Yan et al. (2003) found that over 85% of homoeologous contig pairs showed some level of microsynteny, including 62% that showed extensive microsynteny. Sequences from one contig group showed 94% sequence identity between nine sets of paralogous sequences tested. In further work, Yan et al. (2004) found that 85% of the coding sequences and 75% of noncoding sequences were conserved between duplicate genomic loci in soybean that were examined in detail.

High levels of similarity between duplicated genomic regions have also been found in genomes considered less complicated than soybean. Similarity between homoeologs has been estimated in *A. thaliana*; gene sequences in homoeologous genomic segments showed that 26%–45% of sequences had a high degree of sequence similarity (BLAST scores >150; Blanc et al. 2000). Others have measured divergence of duplicates in *Arabidopsis* that arose in the most recent duplication event, focusing on nonsynonymous and synonymous substitution rates. Ku et al. (2000) found a nonsynonymous substitution rate of 21% divergence between homoeologous regions in *Arabidopsis*. Ziolkowski et al. (2003) measured the rate of synonymous substitutions and found a 79%–88% rate of synonymous substitutions.

Sequence divergence between paralogous sequences in soybean is far less than that in *Arabidopsis*, suggesting that soybean underwent a large-scale duplication event more recent than that of *A. thaliana*. The fact that RFLP probes identified an average of three genomic regions suggests that more than one large-scale duplication event occurred in the evolution of the soybean genome. Indeed, both recent and ancient large-scale duplication events in the soybean genome have been suggested in other studies (Yan et al. 2003, 2004; Lee et al. 2001; Shoemaker et al. 1996). Furthermore, large segmental duplications have been previously identified in soybean, ranging in length up to 100 cM, with an average of 45 cM (Shoemaker et al. 1996).

The wide-scale prevalence of segmentally duplicated regions needs to be considered when comparing genomes. Duplications yielding networks of synteny lead to complex relationships between (and within) genomes. Understanding these networks is critical in successfully leveraging information from one system to another, including model systems.

Work within a single genome is also affected by highly similar duplicated genomic regions as seen in soybean. For example, these regions may make it difficult to assemble DNA fragments or sequence, hinder functional genomics by masking mutant phenotypes, and make chromosome walking difficult. On the other hand, duplicated genomic regions can also be exploited by genomics researchers. An understanding of duplicated genomic regions in *Arabidopsis* is expected to improve gene prediction and annotation by aligning and comparing duplicate genomic regions (Blanc et al. 2000)

Gene organization

A majority of RFLP probes used in this study were derived from *Pst*I digestion of soybean genomic DNA (Keim et al. 1990), a methyl-sensitive enzyme that cuts only in hypomethylated regions of the genome. Because hypomethylated regions of the genome tend to be gene rich (Burr et al. 1988; Grant et al. 2000), the distribution of RFLP probes within the BAC library is expected to reflect the distribution of genes and gives insight into gene clustering in soybean. Grant et al. (2000) found that 72% of RFLPs appeared to be coding sequences based on homology to *A. thaliana* genomic and cDNA sequences. Indeed, RFLP-anchored BAC end sequences were found to have 80% more gene-like sequences and 40% fewer repetitive sequences than simple sequence repeat derived BAC ends, supporting the hypothesis that RFLPs sample from a gene-rich portion of the genome (Marek et al. 2001). Nevertheless, the same study failed to find evidence of gene clustering as measured by clustering of gene-like BAC end sequences within RFLP-anchored BAC contigs. However, it is possible that gene clustering exists but was not detectable on the scale of sampling within BAC contigs (Marek et al. 2001).

Poisson distributions have been used in biology to simulate random sampling and explore clustering. Distributions matching Poisson curves indicate that each event occurs with equal probability or, in other words, that sampling is random and there is no clustering. For instance, Poisson distributions have been used to investigate clustering of amino acids in protein chains (Cai et al. 2002), the distribution of linkage blocks in the human genome (Clark 1999), the distribution of genomic mutations after mutagenesis (Belouchi et al. 1996; Topal et al. 1986), and the distribution of transposable elements within *Drosophila* genomes (Charlesworth et al. 1992; Nuzhdin 1995). Here, we used the same reasoning to investigate physical clustering of genes.

The number of times RFLP probes identified unique BACs versus the number of times they identified a BAC previously found by another probe is significantly different from the expected Poisson distribution if RFLP probes are randomly sampling all of the BACs in the libraries. Conversely, numbers of unique versus redundant BAC hits are not significantly different from a theoretical distribution simulating sampling of just 24% of the BACs in the libraries. This suggests that RFLPs are not sampling the genome randomly but rather are clustered, sampling randomly from only one quarter of the genome. Consequently, RFLPs and, by extension, genes may be clustered in 24% of the genome or less. Indeed, when looking for physically linked RFLP probes, in some cases as many as four linked RFLPs could be assigned to a single coalesced BAC contig, providing further evidence of clustering.

Evidence of gene clustering has been observed in several species in the legume family for which data are available. Pea is estimated to have most of its genes clustered in only 20% of its genome based on probing DNA fractions with gene-like probes (Barakat et al. 1999). The estimate of one gene per 9.9 kb in *L. japonicus* translates into gene-rich portions of the genome covering approximately 50% of the genome if *L. japonicus* is assumed to have a similar total number of genes as *A. thaliana* (25 498 genes; The *Arabidopsis* Genome Initiative 2000). Similarly, the estimate of 6.5 kb per gene in *M. truncatula* translates into an estimate of approximately one third of the genome being gene rich (D. Cook and D.J. Kim, personal communication). Cytogenetic studies have pointed to even a smaller fraction of the *M. truncatula* genome, 20%, containing the majority of genes (Kulikova et al. 2001). Moreover, our previous gene density estimates of one gene per 8 kb in soybean (Young et al. 2003) translate into gene-rich regions of the genome covering less than 20% of the genome, comparable with the results reported here.

Our estimate of gene clustering in 25% or less of the soybean genome is therefore not surprising. With a genome size of 1103 Mb (Bennett and Leitch 2003), gene-rich regions in the soybean genome are predicted to occupy 275 Mb or less. The contigs described in this study fall into these RFLP-rich and likely gene-rich regions because they were identified with RFLP probes. These contigs cover approximately 10% of the genome but presumably more than 40% of gene-rich regions. Additional contigs anchored with RFLPs, not discussed in this study, bring this total up to as much as 50% of the gene-rich regions (J. Mudge et al., unpublished).

Summary

Insights into the soybean genome have been uncovered using networks of BAC contigs from duplicate genome regions identified with RFLP probes. Paralogous sequences from duplicate regions of the soybean genome were highly conserved, sharing 86–100% sequence identity. Duplicated contigs exhibited frequent cases of microsynteny. The use of RFLP probes from hypomethylated and presumably gene-rich regions provided an indication of gene distribution in the soybean genome. RFLP probes and, by extension, genes appear to be clustered in 25% of the genome or less. We therefore predict that most genes occupy approximately 275 Mb of the 1100-Mb soybean genome.

Understanding the structure of duplicated genomic regions and the organization of genes will aid in reconstructing the evolution of the genome and in developing strategies for genetic and genomic studies. Understanding the current genome structure and how it evolved is critical in linking soybean's evolutionary history to that of other species to help leverage information from multiple organisms.

The contig sets identified in this study provide a valuable resource for developing an understanding of duplicate regions. With information from duplicate genomic regions, one can look at paralogous genes that perform the same function or have evolved to perform differing functions (Lynch and Conery 2000; Walsh 1995). On a broader scale, one can look at the evolution of gene families, biological adaptation, and niche radiation, as well as look at general mechanisms of genome evolution (e.g., see Blanc et al. 2000; Bowers et al. 2003; Cannon and Young 2003; Grant et al. 2000; Ku et al. 2000; Zhang et al. 2001).

Resources described in this paper also provide tools to explore the relationship of structural and functional genetic redundancy and its significance in soybean. Only through understanding genome structure, including the organization of genes and the evolution of duplicated regions, can we begin to understand the evolutionary opportunities and challenges that result from segmental duplication.

Acknowledgements

We thank S. Cannon for helpful discussions and R. Staggs, T. Schmidt, C. Paule, and A. Shah for assistance with some computational analysis. We thank D. Foster-Hartnett, A. Cooper, K. Larson, S. Grivna, and D. Gulya for technical assistance. This research was supported by National Science Foundation grants DBI 98-72664 and DBI 98-72565. This paper is published as part of the series of the Minnesota Agricultural Experiment Station.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402.

Arumuganathan, K., and Earle, E.D. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. **9**: 208–219.

Bairoch, A., and Apweiler, R. 2000. The SWISS-PROT protein database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28**: 45–48.

Barakat, A., Carels, N., and Bernardi, G. 1997. The distribution of genes in the genomes of Gramineae. Proc. Natl. Acad. Sci. U.S.A. **94**: 6857–6861.

Barakat, A., Han, D.T., Benslimane, A., Rode, A., and Bernardi, G. 1999. The gene distribution in the genomes of pea, tomato and date palm. FEBS Lett. **463**: 139–142.

Barker, W.C., Garavelli, J.S., Huang, H., McGarvey, P.B., Orcutt, B.C., Srinivasarao, G.Y., Xiao, C., Yeh, L.L., Ledley, R.S., Janda, J.F., Pfeiffer, F., Mewes, H.-W., Tsugita, A., and Wu, C. 2000. The protein information resource (PIR). Nucleic Acids Res. **28**: 41–44.

Belouchi, A., Ouimet, M., Dion, P., Gaudreault, N., and Bradley, W.E. 1996. Influence of alkyltransferase activity and chromosomal locus on mutational hotspots in Chinese hamster ovary cells. Proc. Natl. Acad. Sci. U.S.A. **93**: 121–125.

Bennett, M.D., and Leitch, I.J. 2003. Plant DNA C-values database (release 2.0, January 2003). Available from http://www.rbgkew.org.uk/cval/homepage.html [accessed on 11 March 2004].

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. 2003. GenBank. Nucleic Acids Res. **31**: 23–27.

Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, I. 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell, **12**: 1093–1101.

Blanc, G., Hokamp, K., and Wolfe, K.H. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. **13**: 137–144.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature (Lond.), **422**: 433–438.

Burr, B., Burr, R., Thompson, K., Albertson, M., and Struber, C. 1988. Gene mapping with recombinant inbreds in maize. Genetics, **118**: 519–526.

Cai, Y., Dodson, C.T., Doig, A.J., and Wolkenhauer, O. 2002. Information-theoretic analysis of protein sequences shows that amino acids self-cluster. J. Theor. Biol. **218**: 409–418.

Cannon, S.B., and Young, N.D. 2003. OrthoParaMap: distinguishing orthologs from paralogs by integrative comparative genome data and gene phylogenies. BMC Bioinformatics, **4**: 35.

Carels, N., Barakat, A., and Bernardi, G. 1995. The gene distribution of the maize genome. Proc. Natl. Acad. Sci. U.S.A. **2192**: 11 057 – 11 060.

Charlesworth, B., Lapid, A., and Canada, D. 1992. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster.* I. Element frequencies and distribution. Genet. Res. **60**: 103–114.

Clark, A.G. 1999. The size distribution of homozygous segments in the human genome. Am. J. Hum. Genet. **65**: 1489–1492.

Cregan, P.B., Jarvik, T., Bush, A.L., Shoemaker, R.C., Lark, K.G., Kahler, A.L., Kaya, N., VanToai, T.T., Lohnes, D.G., Chung, J., and Specht, J.E. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. **39**: 1464–1490.

Danesh, D., Peñuela, S., Mudge, J., Denny, R.L., Nordstrom, H., Martinez, J.P., and Young, N.D. 1998. A bacterial artificial chromosome library for soybean and identification of clones near a major cyst nematode resistance gene. Theor. Appl. Genet. **96**: 196–202.

Ewing, B., and Green, P. 1998. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. Genome Res. **8**: 175–185.

Foster-Hartnett, D., Mudge, J., Larsen, D., Danesh, D., Yan, H.H., Denny, R., Peñuela, S., and Young, N.D. 2002. Comparative genomic analysis of sequences sampled from a small region on soybean (*Glycine max*) molecular linkage group G. Genome, **45**: 634–645.

Gaut, B.S. 2001. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. Genome Res. **11**: 55–66.

Gill, K.S., Gill, B.S., Endo, T.R., and Boyko, E.V. 1996*a*. Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. Genetics, **143**: 1001–1012.

Gill, K.S., Gill, B.S., Endo, T.R., and Taylor, T. 1996*b*. Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. Genetics, **144**: 1883–1891.

Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. Japonica). Science (Washington, D.C.), **296**: 92–100.

Grant, D., Cregan, P., and Shoemaker, R.C. 2000. Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc. Natl. Acad. Sci. U.S.A. **97**: 4168–4173.

Kaneko, T., Asamizu, E., Kato, T., Sato, S., Nakamura, Y., and Tabata, S. 2003. Structural analysis of a *Lotus japonicus* genome. III. Sequence features and mapping of sixty-two tac clones which cover the 6.7 Mb regions of the genome. DNA Res. **10**: 27–33.

Keim, P., Diers, B., Olson, T., and Shoemaker, R.C. 1990. RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics, **126**: 735–742.

Ku, H.M., Vision, T., Liu, J.P., and Tanksley, S.D. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. U.S.A. **97**: 9121–9126.

Kulikova, O., Gualtieri, G., Geurts, R., Kim, D.J., Cook, D., Huguet, T., de Jong, J.H., Fransz, P.F., and Bisseling, T. 2001. Integration of the FISH pachytene and genetic maps of *Medicago truncatula*. Plant J. **27**: 49–58.

Lee, J.M., Grant, D., Vallejos, C.E., and Shoemaker, R.C. 2001. Genome organization in dicots. II. *Arabidopsis* as a ‘bridging species’ to resolve genome evolution events among legumes. Theor. Appl. Genet. **103**: 765–773.

Lynch, M., and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science (Washington, D.C.), **290**: 1151–1155.

Marek, L.F., and Shoemaker, R.C. 1997. BAC contig development by fingerprint analysis in soybean. Genome, **40**: 420–427.

Marek, L.F., Mudge, J., Darnielle, L., Grant, D., Hanson, N., Paz, M., Huihuang, Y., Denny, R., Larson, K., Foster-Hartnett, D., Cooper, A., Danesh, D., Larsen, D., Schmidt, T., Staggs, R., Crow, J.A., Retzel, E., Young, N.D., and Shoemaker, R.C. 2001. Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome, **44**: 572–581.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. 1997. High throughput fingerprint analysis of large-insert clones. Genome Res. **7**: 1072–1084.

Nuzhdin, S.V. 1995. The distribution of transposable elements on X chromosomes from a natural population of *Drosophila simulans*. Genet. Res. **66**: 159–166.

Panstruga, R., Buschges, R., Piffanelli, P., and Schulze-Lefert, P. 1998. A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res. **1526**: 1056–1062.

Peñuela, S., Danesh, D., and Young, N.D. 2002. Targeted isolation, sequence analysis, and physical mapping of nonTIR NBS-LRR genes in soybean. Theor. Appl. Genet. **104**: 261–272.

Pruitt, R.E., and Meyerowitz, E.M. 1986. Characterization of the genome of *Arabidopsis thaliana*. J. Mol. Biol. **187**: 169–183.

Shoemaker, R.C., and Olson, T.C. 1993. Molecular linkage map of soybean (*Glycine max* L. Merr.). *In* Genetic maps: locus maps of complex genomes. *Edited by* S.J. O’Brian. Cold Spring Harbor Laboratory Press, Plainview, N.Y. pp. 6131–6138.

Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N.D., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G., and Boerma, H.R. 1996. Genome duplication in soybean (*Glycine* subgenus soja). Genetics, **144**: 329–338.

Shoop, E., Chi, E., Carlis, J., Bieganski, P., Riedl, J., Dalton, N., Newman, T., and Retzel, E. 1995. Implementation and testing of an automated EST processing and analysis system. *In* Proceedings of the 28th Annual Hawaii International Conference on System Sciences, Maui, Hawaii, 3–6 January 1995. Vol. 5. *Edited by* L. Hunter. IEEE Computer Society Press, Los Alamitos, Calif. pp. 52–61.

Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M., and Van de Peer, Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U.S.A. **99**: 13 627 – 13 632.

Soderlund, C., Longden, I., and Mott, R. 1997. FPC — a system for building contigs from restriction fingerprinted clones. Comput. Appl. Biosci. **13**: 523–535.

Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T., and Coulson, A. 1988. Software for genome mapping by fingerprinting techniques. Comput. Appl. Biosci. **4**: 125–132.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature (Lond.), **408**: 796–815.

Topal, M.D., Eadie, J.S., and Conrad, M. 1986. *O*6-methylguanine mutation and repair is nonuniform. Selection for DNA most interactive with *O*6-methylguanine. J. Biol. Chem. **261**: 9879–9885.

Vision, T.J., Brown, D.G., and Tanksley, S.D. 2000. The origins of genomic duplications in *Arabidopsis*. Science (Wash., D.C.), **290**: 2114–2117.

Walsh, J.B. 1995. How often do duplicated genes evolve new functions? Genetics, **139**: 421–428.

Wendel, J.F. 2000. Genome evolution in polyploids. Plant Mol. Biol. **42**: 225–249.

Yan, H.H., Mudge, J., Kim, D.J., Shoemaker, R.C., Cook, D.R., and Young, N.D. 2003. Estimates of conserved microsynteny among the genomes of *Glycine max*, *Medicago truncatula* and *Arabidopsis thaliana*. Theor. Appl. Genet. In press.

Yan, H.H., Mudge, J., Kim, D.-J., Shoemaker, R.C., Cook, D.R., and Young, N.D. 2004. Comparative physical mapping reveals features of microsynteny between *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana*. Genome, **47**: 141–155.

Young, N.D., Mudge, J., and Ellis, T.N. 2003. Legume genomes: more than peas in a pod. Curr. Opin. Plant Biol. **6**: 199–204.

Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science (Washington, D.C.), **296**: 79–92.

Zhang, L., Pond, S.K., and Gaut, B.S. 2001. A survey of the molecular evolutionary dynamics of twenty-five multigene families from four grass taxa. J. Mol. Evol. **52**: 144–156.

Ziolkowski, P.A., Blanc, G., and Sadowski, J. 2003. Structural divergence of chromosomal segments that arose from successive duplication events in the *Arabidopsis* genome. Nucleic Acids Res. **31**: 1339–1350.